*Application for*
*UNITED STATES LETTERS PATENT*

*Of*

TSUNEHIKO WATANABE

JUNJI YOSHII

TADASHI MIZUNUMA

YUICHI MINESAKI

FUMITOSHI OGURA

KEISUKE YAMAMOTO

AND

TATEO NAGAI

*For*

# DATA DISTRIBUTION METHOD, DATA SEARCH METHOD, AND DATA SEARCH SYSTEM

# DATA DISTRIBUTION METHOD, DATA SEARCH METHOD, AND DATA SEARCH SYSTEM

## BACKGROUND OF THE INVENTION

### 1. Technical Field

The present invention relates to a method of retrieving information from a plurality of databases in which information about biological substances, such as sequences of bases and proteins, are stored, by associating the databases with one another and thus clarifying the connections among them.

### 2. Background Art

Databases storing information about biological substances exist all over the world and are available to the public on the Web. Biology researchers can take advantage of those databases for their own studies (see Non-patent document 1). Open databases related to gene information and protein information have their own unique registration numbers (to be hereafter referred to as IDs), which are in many cases assigned to the genes and proteins stored in the databases. So far, when a researcher searches open databases for his own data to retrieve data from the databases, it has been necessary for him to relate his own data with the ID of the particular database using some kind of means. According to the most typical method for that purpose, a homology search is carried out between the base sequence or protein sequence the researcher possesses and the base sequence or protein sequence stored in the database, such that they can be associated with one another.

This can be carried out in two ways. One has the researcher search the open databases on the Web for his own data one-by-one. The other involves the researcher downloading the data of the databases on the Web into his own facility one-by-one and then searching the data, in order to avoid the chances of information leakage that could happen during a search via the Internet. Fig. 21 schematically shows a search system according to the prior art whereby the data

is downloaded from databases on the Web. A user 218 downloads files 219 one by one from an open database 211 via the Internet 212 to a facility 217 of the user. The user 218 then carries out a search on the thus downloaded files 219.

(Non-patent document 1) Baxebanis, A. D: Nucl. Acids Res., 28:1-10, 2000, "Genetics Databases" (Bishop M. J. ed.), Academic Press, Cambridge, 1999.

## SUMMARY OF THE INVENTION

It has been possible to search for information on the Web on a one-by-one basis because of the limited number of data items, typically in the range between one to 10, that had to be handled by the researcher at one time. However, the recent technological advances have allowed hundreds or even thousands of data items to be handled, making it extremely burdensome to search them one by one. The search conducted on a plurality of open databases has also resulted in the creation of unnecessary data, from which the researcher had to re-extract information of his interest. Furthermore, there are so many databases around the world that the researcher has to evaluate and decide on which databases are necessary for him. Some databases contain a plurality of biological species (such as humans, mice, and rice), and no systems have been available for retrieving data concerning a certain living species from various databases in a comprehensible manner. Nor have there been any systems for retrieving data according to the type of data (DNA, mRNA, EST).

In the case of downloading data from open databases one by one into the user's facility, this could take so much time if the amount of data to be downloaded is large that the line could be interrupted in the middle of the downloading operation. Such downloading also requires the line to be occupied for a long time. In addition, the amount of bio-related information is increasing at such a rapid pace that it is expected that the downloading operation would be

more and more time-consuming and complicated. Further, as the information in the open databases are managed by individual database administrators, it has been difficult for the biology researchers to be constantly informed of the update period or the current number of data items in the individual open databases.

There are also various links between databases. Accordingly, data search has been conducted by tracking a plurality of links. For example, as shown in Fig. 22, when obtaining data in a database D that corresponds to data in a database A, there are a route that goes through a database B and another route that goes through a database C. Data in the database B that corresponds to a gene A1 in the database A are B1 and B2, to which data D1 and D2 in the database D correspond. Data in the database C that corresponds to the gene A1 is C1, to which data D3 in the database D corresponds. In this example, there are three items of data D1, D2 and D3 in the database D that correspond to the gene A1 in the database A, so the user has to re-examine which is the correct data.

In light of the above-described problems associated with the search on databases regarding information about biological substances, it is the object of the invention to provide a method and system for enabling data in databases on the network to be easily retrieved.

In accordance with the invention, necessary information is extracted from a plurality of databases to create an index, which is then distributed. Thus, the user can obtain only necessary information. As a plurality of items of data are put together in a single index, the amount of data can be reduced and the download from the data center to the user facility can be smoothly carried out, so that the problem of the line being occupied during download for a long time can be avoided. Furthermore, as the updating of the databases and changes in format, for example, can be effected together at the data center, the user can be spared of bothersome work required for those purposes. In cases where there are no chances of information leakage, for example, the user may directly access an

index placed at the data center and conduct a search without downloading it into the user facility.

The invention provides a data distribution method comprising the steps of: downloading data from a plurality of databases in which information about biological substances is stored; extracting from the downloaded data information indicating a link between data in two databases, a detailed description of each data, and sequence data for homology search, which together constitute an index; and distributing the thus extracted index.

The invention also provides a data search method comprising the steps of: downloading data from a plurality of databases in which information about biological substances is stored; extracting from the downloaded data information indicating links between data in two databases; receiving a start database name, a target database name, and a data ID in the start database, which together constitute a search key; acquiring a data ID of the target database by following those links among the extracted links between data that match the predetermined order of the link between a plurality of databases, while referring to information indicating the predetermined order of the link between the databases and using the received data ID in the start database as a start point; and displaying the thus acquired data ID of the target database.

The invention further provides a data search method comprising the steps of: downloading data from a plurality of databases in which information about biological substances is stored; extracting from the downloaded data information indicating links between data in two databases and sequence data for homology search; receiving a start database name, a target database name, and input sequence data, which together constitute a search key; conducting a homology search for homology-search sequence data in the start database, using the input sequence data; acquiring a corresponding data ID of the target database by following those links among the extracted links between data that match the predetermined order of a link between databases, while referring to information

indicating the predetermined order of the link between the databases and using as a start point the data ID in the start database that has been acquired by the homology search; and displaying the thus acquired data ID of the target database.

The invention further provides a data search method comprising the steps of: preparing index data that is a collection of information indicating links between data in two databases, based on a plurality of databases in which information about biological substances is stored; preparing a table defining the order of the links between the plurality of databases; receiving a start database name, a target database name, and a data ID of the start database, which together constitute a search key; acquiring a corresponding data ID in the target database by following those links among the links between data that match the order of the links between the databases, while using as a start point the data ID in the start database that has been received; and displaying the acquired data ID of the target database.

The invention further provides a data search method comprising the steps of: preparing index data that is a collection of information indicating links between data in two databases and sequence data for homology search, based on a plurality of databases in which information about biological substances is stored; preparing a table defining the order of links between the plurality of databases; receiving a start database name, a target database name, and input sequence data, which together constitute a search key; conducting a homology search for homology-search sequence data in the start database, using the input sequence data; acquiring a corresponding data ID of the target database by following those links among the links between the data that match the order of the links between the plurality of databases, using as a start point the data ID in the start database that has been acquired by the homology search; and displaying the acquired data ID of the target database.

The invention further provides a data search system comprising: index data that is a collection of information indicating links between data in two

databases that is gathered from a plurality of databases in which information about biological substances is stored; a table defining the order of the links between the plurality of databases; an input portion for receiving a start database name, a target database name, and a data ID in the start database, which together constitute a search key; a search portion for acquiring a corresponding data ID of the target database by following those links among the links between data that match the order of the links between the databases, while using as a start point the data ID in the start database that has been received; and a display portion for displaying the acquired data ID of the target database.

The invention moreover provides a data search system comprising: index data that is a collection of sequence data for homology search and information indicating links between data in two databases that is gathered from a plurality of databases in which information about biological substances is stored; a table defining the order of the links between the plurality of databases; an input portion for receiving a start database name, a target database name, and an input sequence data in the start database, which together constitute a search key; a first search portion for conducting a homology search for homology-search sequence data in the start database, using the input sequence data; a second search portion for acquiring a corresponding data ID of the target database by following those links among the links between data that match the order of the links between the plurality of databases, using as a start point the data ID in the start database that has been acquired by homology search; and a display portion for displaying the acquired data ID of the target database.

In accordance with the inv*, a search can be conducted on thousands of data items against an index all at once. Further, by classifying and arranging the databases with which a network is constructed by living species (humans, mice, rice, for example) and by the type of data (DNA, mRNA, EST), the user can obtain data matched with his purposes. By preparing a table or the like defining the order of links among a plurality of databases, and by following the links

between the plurality of databases according to the defined route, search result with a reduced amount of noise can be obtained.

## BRIEF DESCRIPTION OF THE DRAWINGS

Fig. 1 is a conceptual chart showing an example of the structure of the biological substance information search system according to the invention.

Fig. 2 shows a flowchart of the procedure for creating an index for retrieving information, based on a plurality of database in which biological substance information is stored.

Fig. 3 shows a procedure for creating link information.

Fig. 4 shows another example of routes obtained from the link information.

Fig. 5 shows an example of a route table in which information about routes (order) of links among databases is stored.

Fig. 6 shows an example of a network display of the contents of the route table.

Fig. 7 illustrates the effect of limiting the links among databases.

Fig. 8 shows a procedure for creating homology search data.

Fig. 9 shows a procedure for creating a detailed description file.

Fig. 10 shows the details of index information.

Fig. 11 shows a flowchart of the procedure for biological substance information search according to the invention.

Fig. 12 shows a block diagram of the search system according to the invention.

Fig. 13 shows an example of an interface used when searching for the ID of a database.

Fig. 14 shows an example of an interface used when searching for a sequence.

Fig. 15 shows an example of input data.

Fig. 16 shows an example of the screen of a display portion on which search result is displayed.

Fig. 17 shows an example of display of detailed description.

Fig. 18 shows an example of an input data file.

Fig. 19 shows an example of the screen of a display portion on which search result is displayed.

Fig. 20 shows an example of display of homology search result.

Fig. 21 schematically shows a conventional search system that downloads data from databases on the Web.

Fig. 22 shows an example of conducting a search by following a plurality of links.

## DESCRIPTION OF THE PREFERRED EMBODIMENTS

The invention will be hereafter described by way of embodiments with reference made to the drawings.

Fig. 1 is a conceptual chart of an example of the structure of the biological substance information search system according to the invention. Data in an open or commercial database 11 is downloaded to a data center 13 via the Internet 12. In the data center 13, an index 15 is created based on the downloaded data. The thus created index 15 is delivered (index 16) to a facility 17 of the user 18, who conducts a search on the index 16.

The index includes link information indicating the correspondence among data contained in different databases, detailed description of each data, and homology-search data. The detailed description of each data refers to the detailed description of entries stored in each entry in the database. The homology-search data refers to information about sequences such as base sequence or protein sequence contained in the database. The user conducts a homology search between the base sequence or protein sequence he possesses and the base sequence or protein sequence in the data of a target open database. The

homology search usually employs software called BLAST. Thus, for the data subjected to homology search, a FASTA-format sequence data is usually formatted for BLAST.

By classifying and organizing the databases employed in constructing the network by the living species (humans, mice and rice, for example) and by the type of data (DNA, mRNA, EST), the user can obtain data according to his purposes.

Fig. 2 shows a flowchart of the procedure for creating the index for information search, based on a plurality of databases storing information about biological substances.

In step 11, data is downloaded from the open databases, such as official databases and commercial databases, to the data center. In step 12, the link information, homology-search data, and the detailed description of each ID are automatically extracted from the downloaded data. The homology search data is obtained from all databases to be registered in the index in which sequence information exists. The detailed information is obtained from all of the databases to be registered in the index. Finally, in step 13, the link information, homology search data and the detailed description of each ID are together delivered to the user's facility.

Fig. 3 illustrates the procedure for creating the link information in step 12 of Fig. 2. In the illustrated example, a database A corresponds to databases B and E such that an entry A1 in the database A corresponds to an entry B1 in the database B and an entry E1 in the database E. These correspondences are described in a database A file. Thus, individual IDs are taken out of the database A file, and A1 in the database A and B1 in the database B are stored in a table 31. Similarly, there is described the correspondence between the entry A1 in the database A and the entry E1 in the database E, and therefore this correspondence is stored in a table 32. A database B file describes the correspondence between the entry B1 in the database B and an entry C1 in a

database C, which are taken out and stored in a table 33. A database C file describes the correspondence between the entry C1 in the database C and an entry D1 in a database D, which are taken out and stored in a table 34. By combining these tables 31, 33 and 34, a table 35 can be created. The tables 32 and 35 can be schematically described by way of a link chart shown in Fig. 36.

Fig. 4 shows another example of the route obtained from the link information. In the tables stored in the databases as link information, the IDs of two databases are associated with each other, as shown in the tables 31 to 34 in Fig. 3. Based on these tables, a table 41 or 42 is created, as shown in Fig. 4. The relationship between these tables can be described by way of a link chart 43. By tracing the corresponding data on the link chart 43, the data D1 in the database D that corresponds to the data A1 in the database A, for example, can be retrieved.

The databases contain link information linking them to other various databases. As a result, the problem described above with reference to Fig. 22 could occur due to complications among the links. Thus in the present invention, the links among the databases are limited such that the individual databases are linked to one another according to a predetermined rule (order). This limitation imposed on the links among the databases will be described below.

Fig. 5 shows an example of a route table in which information concerning the permitted routes (order) among the databases is stored. "KeyDB" designates a database as a search starting point. "TargetDB" designates a database to be searched for data corresponding to data in KeyDB. The databases A, B, C, and so on including open databases, commercial databases and private databases, may in some cases describe inter-data linkage information indicating to which data in other databases certain data in a particular database corresponds. In such cases, it is possible to follow various routes to search for the data in TargetDB that corresponds to designated data in KeyDB. However, if all of the link information is to be utilized, there is a chance of picking up noise information, as

mentioned above. Therefore, the route (order) of the link from KeyDB to TargetDB, once KeyDB and TargetDB are designated, is uniquely designated by the route table. In the illustrated example, in the case where KeyDB is A and TargetDB is C, the data in the database C that corresponds to data in the database A is retrieved by following the link in the order of the database A, database B and database C, referring to the route table of Fig. 5. Similarly, when KeyDB is A and TargetDB is D, the data in the database D that corresponds to data in the database A is retrieved by following the link in the order of the database A, database B, database C and database D, while referring to the route table of Fig. 5.

Fig. 6 shows an example of the contents of the route table by way of a network. The fact that databases 61 and 63 correspond to each other is indicated by a line 62 connecting these two databases. It is now assumed that the database 61 has been newly created on the basis of the data stored in the database 63, and that data A stored in the database 61 corresponds to data B stored in the database 63. In the present invention, only such link information that follows the origin of the data is utilized, and even if link information is stored in the database 61 that is related to another database 64, such information is not utilized as the link information for retrieval. By thus limiting the link between databases, the acquisition of unnecessary data can be limited.

Fig. 7 shows the effect of limiting the link between databases, the figure corresponding to Fig. 22.

In the case where the database A describes link information to the database B and link information to the database C, the present invention utilizes only the link information between the database A and the database C, which is more reliable, and does not utilize the link information between the database A and the database B. As a result, gene data D3 in the database D that corresponds to gene data A1 in the database A can be acquired. Thus, by limiting the link between the databases, the acquisition of unwanted data that produces noise, as

described with reference to Fig. 22, can be limited, so that only appropriate data can be acquired.

Fig. 8 shows the procedure for creating the homology search data in step 12 of Fig. 2. In the illustrated example, an ID 83 and sequence data 84 for each entry are extracted from a file 81 downloaded from an open database, and a file 82 in which sequence data 85 of FASTA format is stored is created.

Fig. 9 shows the procedure for creating a detailed-description file in step 12 of Fig. 2. In the illustrated example, an ID 93 for each entry and a detailed description 94 concerning the entry are extracted from a file 91 downloaded from an open database, and they are then stored in a detailed-description file 92 as a pair 95 of the ID and the detailed description.

Fig. 10 shows the details of the index information. The index information (link information 101, detailed description 103, and homology search data 106) is created in the data center 13. The link information 101 is retained in the form of link tables 102. The detailed description 103 is retained in the form of detailed description tables 104. The individual tables of the link information and detailed description are stored in a database 107. A file 105 of FASTA format is formatted for BLAST, thereby creating homology search data 106. The index information thus created in the data center 13 is transferred to a user facility 17. In this case, a copy of the database 107 is created in a database 108 in the user facility 17 by a replication process. Further, a copy 109 of the homology search data 106 is transferred to the user facility 17. Also, a copy 111 of a route table 110 in which information about the route (order) of the links among the databases is stored is transferred to the user facility 17.

Fig. 11 shows a flowchart of the procedure for biological substance information search according to the invention. Fig. 12 shows a block diagram of a search system for realizing this search method.

The search system of the invention includes a database 124 in which the link information and detailed description described with reference to Fig. 10 are

stored, homology search data 125, a route table 126 describing the order of links among databases, an input operation portion 127, a display portion 128 for displaying search results, and a search processing portion 121. The search processing portion 121 includes an ID search portion 122 for conducting ID search by following the links, and a homology search portion 123 for conducting homology search between sequence data inputted from the input operation portion and homology search data. Figs. 13 and 14 illustrate examples of an input interface used during data search. Fig. 13 shows the input interface used when searching for the ID of a database, and Fig. 14 shows the input interface used when searching for a base sequence or a protein sequence.

First, the search method and system whereby the ID of user data is converted into the ID of a database on the network will be described.

In step 21 of Fig. 11, the input operation portion 127 is operated to input data. For example, as an input-data file as shown in Fig. 15 is selected by a "File Upload" button 132 shown on the screen of Fig. 13, the data is displayed in a data input field 131 shown in Fig. 13, separated by commas. The input data can be cleared by depressing a "Clear" button 133. The input data example shown in Fig. 15 is UniGene data publicly available from NCBI.

In step 22 of Fig. 11, KeyDB and TargetDB are set. A database with the same ID as that of the input data is selected from a KeyDB list 134 shown in Fig. 13, and a database as the object of conversion is selected from a TargetDB list 135 of Fig. 13. Then, a search route is displayed in a field 136 by referring to the route table 126. As a button 137 is selected, the entire view of the ID network as shown in Fig. 6 is displayed, where the KeyDB and TargetDB can be confirmed.

Then, in step 23, a search start button 138 is depressed to start a search. A search program in the ID search portion 122 follows the designated search route to search for a data ID of TargetDB that corresponds to the data ID of KeyDB that has been inputted.

The routine then proceeds to step 24, in which the search result is displayed. Fig. 16 shows an example of the display screen of the display portion 128 on which the search result is displayed. In the illustrated example, entries 163 in SWISS-PROT, which is Target DB, that correspond to entries 162 in UniGene, which is KeyDB, are shown in a field 161. In "Hit Count" 166, the number of the entries 163 in TargetDB that correspond to the entries 162 in KeyDB is shown. By clicking a KeyDB button or TargetDB button 164, a detailed description shown in Fig 17 is displayed. By clicking a "View Route" button 165, a chart showing the search route among databases as shown in Fig. 6 is displayed.

Next, an example where a base sequence or protein sequence the user wishes to search for is converted into the ID of a database on the ID network will be described.

In step 21 of Fig. 11, the sequence data to be retrieved is inputted via the input operation portion 127. For example, as a "File Upload" button 146 on the input screen of Fig. 14 is clicked and an input data file shown in Fig. 18 is selected, the input sequence data is displayed in a data input field 141 of the input screen. By clicking the "Clear" button, the data input field 141 can be emptied.

The routine then advances to step 22, where KeyDB and TargetDB are set. A database (KeyDB) that is desired to be associated with the search data is selected from a DB list 149 in the input screen shown in Fig. 14, and a target database (TargetDB) as the object of conversion is selected from a TargetDB list 143 of Fig. 14. After KeyDB is set, an appropriate BLAST technique is selected from a program list 142, depending on whether the sequence data to be retrieved and the data stored in the database as KeyDB are nucleotide sequences or protein sequences. For example, "blastn (DNA Query vs. DNA DB)" searches for the data of the input nucleotide sequence in the nucleotidesequence database. "blastp (Protein Query vs. Protein DB)" searches for the data of the input protein sequence in the protein sequence database. "blastx (DNA Query vs. Protein

DB)" searches for the data of the input nucleotide sequence in the protein sequence database by performing six-frame translation of the input nucleotide sequence. "tblastn (Protein Query vs. DNA DB)" searches for the data of the input protein sequence in the nucleotide sequence database by performing six-frame translation of a nucleotide sequence database dynamically. Detailed parameters for BLAST search are set in a details option setting portion 147.

As a "View Route" button 144 is depressed, Fig. 6, which is the overall view of the database network, is displayed, enabling the locations of KeyDB and TargetDB to be confirmed. In a field 148, the search route set in the route table is displayed.

In step 23, as a search start button 145 is depressed, a search begins. Initially, a search program (BLAST) in the homology search portion 123 is activated, and then a homology search is conducted between the inputted sequence data and the homology search data in the database designated as KeyDB in order to acquire the ID of the candidate data. Then, a search program in the ID search portion 122 is activated, and, using as a starting point the ID of KeyDB obtained by the homology search, a search is conducted for a corresponding ID of TargetDB by following the route of the links set in the route table.

In step 24, the search result is displayed. Fig. 19 shows an example of the screen of the display portion on which the search result is displayed. In the illustrated example, an ID 193 of TargetDB (SWISS-PROT) that corresponds to the ID 192 of KeyDB (Nucleotide (EST)) is shown in a field 191. In "Hit Count" 197, the number of IDs in SWISS-PROT, which is TargetDB, that corresponds to the ID of KeyDB, that is Nucleotide (EST), is shown. By clicking a "KeyDB" button or "TargetDB" button 194, a detailed description as shown in Fig. 17 can be displayed. By clicking a "ViewAlignment" button 195, a homology search result as shown in Fig. 20 can be displayed. In Fig. 20, "E-value" refers to an expected value, and "Score" refers to a homology value (Altschul, S.F., Gish, W., Miller, W., Myers, E.W. & Lipman, D.J. (1990) "Basic

local alignment search tool." J. Mol. Biol. 215: 403-410). An ID search is conducted using the ID with the highest Score data as a search key.

Thus, in accordance with the invention, all of the databases on a network can be easily searched for data by following links in the network.